SURVEY AND SUMMARY

Exon–intron organization of genes in the slime mold *Physarum polycephalum*

Joanna Trzcinska-Danielewicz and Jan Fronk*

Institute of Biochemistry, Warsaw University, Zwirki i Wigury 93, 02089 Warsaw, Poland

Received May 26, 2000; Revised and Accepted July 28, 2000

ABSTRACT

The slime mold *Physarum polycephalum* is a morphologically simple organism with a large and complex genome. The exon–intron organization of its genes exhibits features typical for protists and fungi as well as those characteristic for the evolutionarily more advanced species. This indicates that both the taxonomic position as well as the size of the genome shape the exon–intron organization of an organism. The average gene has 3.7 introns which are on average 138 bp, with a rather narrow size distribution. Introns are enriched in AT base pairs by 13% relative to exons. The consensus sequences at exon–intron boundaries resemble those found for other species, with minor differences between short and long introns. A unique feature of *P. polycephalum* introns is the strong preference for pyrimidines in the coding strand throughout their length, without a particular enrichment at the 3'-ends.

INTRODUCTION

The evolutionary origin(s) of spliceosomal introns, present in the vast majority of eukaryotic genes, remains a mystery. More than 99% of these introns have the highly conserved 5'-GT/AG-3' termini and most probably are removed by a common splicing mechanism with the participation of U2 snRNA (1). This indicates common evolutionary origin of these introns. A very minor group of introns seems to be removed by an alternative, U12 snRNA-dependent, route (2).

While the overall mechanism of spliceosomal intron removal and the terminal dinucleotide sequences are conserved among all eukaryotes, the details of exon–intron organization and additional sequence information contained within introns vary widely among various taxonomic groups (3–8). This suggests that the mechanisms of intron recognition, as opposed to the splicing itself, are also diverse. In a number of cases intron features limited to a particular group of organisms have been shown to be required for correct and efficient splicing. Examples include the AU-rich sequences in dicotyledonous (9), but not in monocotyledonous (10) plants, and G triplets in short mammalian introns (11). The pyrimidine-rich 3'

sequence is dispensable in short, but not in long, *Drosophila* introns (12,13). Relative intron–exon lengths are crucial for correct splicing in both *Drosophila* (12,13) and mammals (14).

The experimental studies on intron recognition provide an ultimate proof of the importance or otherwise of a particular feature. However, the much simpler statistical analyses offer a good indication of the likelihood of a feature being relevant to intron recognition and/or removal, and also help understand the evolution of these processes (15). Such studies need not be limited to the most thoroughly studied model organisms. Indeed, the broader the scope of such analyses, the more likely they are to uncover the evolutionary paths taken by introns and the splicing machinery.

In this paper we present a survey of exon–intron organization in the Myxomycete (slime mold) *Physarum polycephalum*, a morphologically simple organism whose taxonomical position, especially its relation to the 'cellular slime mold' *Dictyostelium discoideum*, is a matter of some controversy (16,17). *Physarum polycephalum* has a very large genome [$\sim 3 \times 10^8$ bp (18)], larger than those of the evolutionarily far more advanced metazoans *Drosophila melanogaster* and *Caenorhabditis elegans* and the dicotyledonous plant *Arabidopsis thaliana*. This high DNA content is at least in part due to the abundant presence of simple and inverted repeat sequences, some of them transposon-like (19). In this respect *P. polycephalum* is akin to most vascular plants and vertebrates. It should be interesting to learn whether the exon–intron organization of *P. polycephalum* genes is more similar to that of the 'simple' eukaryotes, to which this organism is more closely related, or to that of the highly advanced ones, which it resembles in its genome size and organization.

MATERIALS AND METHODS

GenBank was searched for *P. polycephalum* genomic protein coding sequences. Fourteen complete genes and five partial sequences were found, containing 64 introns in all. Gene name, as given in the database, is followed by accession number. Where needed, a short definition of the coded product is given in parentheses. *Actin*, X07792; *ardA*, M15272 (actin); *ardB*, X60788 (actin); *ardD*, M59234 (actin); *H41*, X15142 (histone H4); *H42*, X15141 (histone H4); *hapP1*, L39871 (a differentiation-specific protein); *Pprap1*, AF190053 and AF000653; *Pprap1*, U10905; *Pprap2*, AF001390 and AF189162; *proA*, M38037

*To whom correspondence should be addressed. Tel: +48 22 8232046; Fax: +48 22 8232046; Email: fronk@geo.uw.edu.pl

Figure 1. Distribution of (A) intron and (B) internal exon sizes in *Ppolycephalum* genes. Homologous introns of similar length were included only once. Homologous exons were included only once. Note the change of bin size in (A) for introns larger than 200 bp.

(profilin, a differentiation-specific actin binding protein); *proP*, M38038 (profilin, a differentiation-specific actin binding protein); *E-α-tubulin*, X14213; *γ-tubulin*, Y09215; *betB1*, X12371, partial (β-tubulin); *P46*, M89947, partial (a differentiation-specific gene); *redA*, Y18122, partial (a differentiation-specific gene); *spherulin*, Z50151, partial (a differentiation-specific gene); *tef1*, AF016243, partial (protein synthesis elongation factor 1-α). Depending on the type of analysis, the dataset was purged to avoid overrepresentation of homologous sequences. The details are given in the text and figure legends.

RESULTS

Exon-intron organization

The distribution of exon and intron sizes is shown in Figure 1. For the majority of *P.polycephalum* genes the exact boundaries of the transcription unit (i.e., the 5'-end of the first exon and the 3'-end of the last exon) have not been convincingly determined, therefore for all analyses we assume the limits of the transcription unit to coincide with the limits of the protein coding sequence (CDS). This obviously leads to an underestimation of the lengths of the outer exons. For this reason we only show the length distribution for the inner exons, the exact lengths of which are known. These are on average 165 ± 85 bp long (range 11–362; median 146). When the 5'-most and 3'-most exons (defined as above) are also included, the average length of all exons becomes 162 ± 138 bp (range 10–1046). The 5'-most exons are on average 89 ± 49 bp (range 10–225) and the 3'-most ones 229 ± 264 bp (range 36–1046).

Figure 2. Distribution of introns along the protein coding sequences. Each coding sequence was normalized to 100%. Only complete genes were analyzed. Homologous introns of the same phase are shown only once. Black triangles, phase zero introns; grey triangles, phase one; white triangles, phase two.

The average length of an intron is 138 ± 103 bp (range 46–575; median 98). Only 18% of them are longer than 200 bp. The average number of introns per gene is 3.7 (range 1–7; all *P.polycephalum* genomic protein coding sequences in the database contain at least one intron). The average intron density (calculated for complete genes only) is 4.9 per 1000 bp coding sequence, with a rather wide range of 2.2 (the *γ-tubulin* gene, the longest gene in the set, containing only three introns) to 8.8 (*Pprasi*). The two shortest pairs of genes, those for histone H4 (*H41* and *H42*) and those for profilins A and P contain one and two introns, respectively; the highest number of introns, seven, is present in the second-longest gene (for E-α-tubulin).

All introns but one (11 of *proA*, just 5' to the ATG codon) lie within the protein coding part of the genes. It is possible that this finding does not faithfully reflect reality: introns present in non-coding regions are more difficult to identify than those interrupting the CDS. The distribution of introns within the CDS is somewhat skewed towards the 5'-end, with 42% of introns located in the 5'-most quarter and only 11% in the 3'-most one (Fig. 2). In Figure 2 we also indicate the phases of the introns. In the whole set, including also incomplete genes, and counting homologous introns of the same phase only once, 51% of introns are phase zero (between codons), 39% phase one (between first and second nucleotides in a codon), and 10% phase two.

We next examined the nucleotide sequences at the exon-intron boundaries. Since there are indications from other organisms that 'short' and 'long' introns may require different signals for correct recognition, we also conducted this analysis for two arbitrarily defined intron subsets: the short ones (shorter than 70 bp, 14 introns), and the long ones (longer than 200 bp, 11 introns). Table 1 presents the combined results, with the sequences for human genes shown for comparison.

In addition to the differences in boundary sequences evident in Table 1, the short and long introns differ in other aspects as well. Of the 11 long introns three are of unknown location in the incomplete CDSs; of the remaining eight, two are the 3'-most ones and neither is the 5'-most one. The location of all 12 short introns (there are two pairs of homologous introns in the original set of 14 short ones; for the present analysis each homologous pair is counted once) is known; two are the 3'-most ones and five are the 5'-most ones. The long introns tend to lie between long exons (they are on average 194 bp when counting only internal exons of exactly defined length, or at least 200 bp when all exons are taken into account, including those of unknown length), with the upstream exons shorter (175 or at least 189 bp, as above) than the downstream ones (219 or at least 211 bp). Conversely, the short introns are surrounded by shorter exons (average for the 'internal' exons

Table 1. Nucleotide sequences at exon–intron borders

		exon					intron																									exon				
		-5	-4	-3	-2	-1	1	2	3	4	5	6	...	-16	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	1	2	3	4	5		
%G	all	14	24	13	11	68	100	-	5	5	62	2	...	2	3	-	8	21	13	8	10	11	13	17	5	17	5	-	100	54	10	21	21	16		
	<70 bp	7	36	-	21	57	100	-	-	-	79	-	...	-	7	-	7	14	29	14	14	14	21	43	-	29	-	-	100	29	7	21	29	21		
	>200 bp	36	9	9	-	82	100	-	-	-	45	-	...	9	-	-	9	18	-	9	-	9	18	-	9	18	-	-	100	73	9	-	9	9		
%A	all	22	22	48	48	8	-	-	84	14	8	12	...	36	32	36	36	30	25	22	17	24	21	17	13	24	16	100	-	27	25	35	21	27		
	<70 bp	29	-	71	36	29	-	-	100	7	14	7	...	57	14	50	43	36	21	43	29	14	29	29	14	21	29	100	-	21	14	29	21	14		
	>200 bp	18	9	36	73	-	-	-	82	9	18	-	...	36	45	45	45	45	45	9	9	36	9	27	18	18	18	100	-	27	27	36	36	9		
%T	all	19	19	13	25	6	-	100	11	60	16	64	...	36	49	52	40	38	30	49	52	43	48	40	59	41	40	-	-	14	43	17	29	17		
	<70 bp	21	29	14	36	-	-	100	-	71	7	79	...	14	71	43	43	43	36	43	50	43	29	14	64	36	43	-	-	36	50	36	21	43		
	>200 bp	27	27	9	18	18	-	100	18	64	18	82	...	36	36	36	36	27	27	55	45	36	55	36	55	45	45	-	-	-	27	18	18	18		
%C	all	44	35	27	16	17	-	-	-	21	14	22	...	25	16	11	16	11	32	21	21	22	19	25	24	17	40	-	-	5	22	27	30	40		
	<70 bp	43	36	14	7	14	-	-	-	21	-	14	...	29	7	7	7	14	-	7	29	21	14	21	14	29	-	-	-	14	29	14	29	21		
	>200 bp	18	55	45	9	-	-	-	-	27	18	18	...	18	18	18	9	9	27	27	45	18	18	36	18	18	36	-	-	-	36	45	36	64		
%Y	all	63	54	40	41	23	-	100	11	81	30	86	...	61	65	63	56	49	62	70	73	65	67	65	83	58	80	-	-	19	65	44	59	57		
	<70 bp	64	65	28	43	14	-	100	-	92	7	93	...	43	78	50	50	50	50	43	57	72	50	28	85	50	72	-	-	50	79	50	50	64		
	>200 bp	45	82	54	27	18	-	100	18	91	36	100	...	54	54	54	45	36	54	82	90	54	73	72	73	63	81	-	-	-	63	63	54	82		
%W	all	41	41	61	73	14	-	100	95	74	24	76	...	72	81	88	76	68	55	71	69	67	69	57	72	65	56	100	-	41	68	52	50	44		
	<70 bp	50	29	85	72	29	-	100	100	78	21	86	...	71	85	93	86	79	57	86	79	57	58	43	78	57	72	100	-	57	64	65	42	57		
	>200 bp	45	36	45	91	18	-	100	100	73	36	82	...	72	81	81	81	72	72	64	54	72	64	63	73	63	63	100	-	27	54	54	54	27		
human	%G	22	25	18	13	80	100	-	46	13	81	23	...	15	14	13	13	12	11	12	12	13	9	7	7	23	-	-	100	53	22	26	24	20		
	%A	27	27	32	61	9	-	-	49	69	7	16	...	13	11	9	9	8	7	7	7	9	9	7	7	22	4	100	-	23	22	23	22	25		
	%T	24	19	11	13	8	-	100	3	9	6	45	...	39	41	42	44	46	50	45	45	39	41	42	48	22	22	-	-	10	35	25	25	26		
	%C	27	29	39	13	3	-	-	3	8	6	17	...	33	34	35	34	35	32	35	35	39	40	44	38	32	74	-	-	14	21	26	29	29		

Percentage frequency of each nucleotide at a particular position is given. All 64 introns were analyzed, since even the apparently homologous ones have significantly different border sequences. Separate numbers are shown for all introns, those shorter than 70 bp, and those longer than 200 bp. Y, C or T; W, A or T. For comparison, data for human genes are shown, based on over 12 000 sequences excluding non-GT/AG introns (<http://www.introns.com/phylo.phtml?phy=homo>).

is 126 bp, for all exons, including those assumed to begin with the AUG codon or end with a stop codon, respectively, it is 114 bp); again, the upstream exons are shorter (120 or 95 bp) than the downstream ones (144 or 133 bp).

Even a cursory analysis of the base composition of individual *P.polycephalum* introns indicates an asymmetrical distribution of purines and pyrimidines, the latter being significantly overrepresented in the coding strand. Another type of unbalanced intron base composition has been observed in some organisms, the introns being enriched in AT base pairs relative to exons (3,9,13). Prompted by these observations we performed a systematic analysis of all the sequences in the dataset, divided into introns, exons (with the outer exons defined as before) and intergenic sequences (defined as all sequences outside the boundaries of the CDSs). The results are presented in Table 2. As before, short and long introns were extracted for separate analysis.

Table 2. Nucleotide composition of various categories of sequences in the *P.polycephalum* genome

	%AT	%CT, coding strand
All introns	61 (41–80)	65 (39–91)
Short introns	65 (53–80)	62 (44–67)
Long introns	61 (44–75)	65 (43–83)
Exons	48	50
Intergenic	57	49
All sequences	54	54

Values in parentheses show ranges. See text for details.

The three most asymmetric introns are: 91% pyrimidines, 97 bp; 89%, 177 bp; 86%, 71 bp, respectively. All these introns are the 3'-most in their respective genes. Only four introns are <50% pyrimidines (39%, 112 bp; 43%, 432 bp; 44%, 55 bp; 48%, 46 bp); they occupy different positions within the CDSs (third out of four, second out of two, third out of five, and second out of three, respectively).

DISCUSSION

Physarum polycephalum is a morphologically simple organism with a large and complex genome. The exon–intron organization of its genes, examined in this paper, exhibits features apparently unique to this organism, features typical for the evolutionarily old (and usually small-genomed) taxa, as well as those characteristic for the more advanced (and large-genomed) ones. Some of these features seem to be more strongly pronounced in *P.polycephalum* than in any other organism examined so far.

Introns are common in *P.polycephalum* genes; in fact, we found no intronless protein coding genomic sequence in the database. In this respect *P.polycephalum* resembles vertebrates and is markedly different from its closer relatives, where a significant percentage of genes have no introns. Even in the more advanced species intronless genes are not unusual, e.g., in the completely sequenced chromosome 4 of *A.thaliana* 19% of genes have no introns (20); among the 218 genes in the completely sequenced *Adh* region on *D.melanogaster* chromosome 2 (21) we found 30 (14%), at least some of which are verified genes, without introns. On the other hand, in the completely sequenced genome of *C.elegans* no more than 3% of the protein coding genes contain no introns (J.Spieth, personal communication). On average, 3.7 introns interrupt a *P.polycephalum* gene (4.9 per 1 kb of CDS). [A slightly

different intron density, ~4.5 per 1 kb of CDS, can be inferred from figure 1 in Logsdon (22). The author gives no details of the calculation so it is difficult to account for this minor discrepancy.] This is significantly more than in the Apicomplexa, *D.melanogaster*, and most fungi (8,15,22,23).

The overall content of intronic sequences and individual intron size have been found to be (weakly) correlated with genome size (8,24). However, it should be borne in mind that for the majority of species analyzed there is a strong correlation between genome size and the complexity of an organism; consequently, it is not entirely clear that the apparent intron size/genome size correlation among distantly related species has at its basis the size of the genome rather than the taxonomic position. There are, however, examples of pairs of species of very similar complexity, but significantly different genome sizes, which give credence to the claim that it is the genome size rather than the organism's complexity that is, to a varying degree, reflected in the size of introns.

In this context one can mention the vertebrates *Homo sapiens* and *Fugu rubripes*, and the vascular plants *A.thaliana* and *Zea mays*. In the former pair the genome sizes differ by a factor of around eight, and a similar difference is found among a sample of orthologous introns (25). A more systematic study found a much smaller difference in intron sizes between these species (24). It ought to be mentioned, however, that the data used for this analysis deviate substantially from those commonly found in the literature [e.g., the average intron size for human is taken as ~380, compared to 3413 bp in Deutsch and Long (8)]. In the case of the two plants, their genome sizes differ by a factor of ~18, while the difference between their introns is barely noticeable. The average size of introns is 239.7 versus 327.5 bp, the dominant size of introns (our visual inspection of size distribution graphs) is ~87 versus ~110 bp, and the total content of introns per gene is 861.6 versus 964 bp (8).

On the other hand, among three organisms of quite similarly sized genomes (*C.elegans*, *D.melanogaster* and *A. thaliana*), but hugely different complexity (*C.elegans* versus *D.melanogaster* and *A.thaliana*) and very distant evolutionary relationship (*A.thaliana* versus *C.elegans* and *D.melanogaster*) the differences in gene organization are substantial [the reader is referred to Deutsch and Long (8) for details].

How does the data obtained for *P.polycephalum* fit into this rather complex picture? The average intron size (138 bp) with the peak in distribution at ~95 bp resemble most closely those of *A.thaliana*. *Physarum polycephalum* introns are similar to those in fungi [*Schizosaccharomyces pombe*, *Aspergillus nidulans* (8,15) and *Neurospora crassa* (23)] in that they do not exceed several hundred base pairs, but, unlike in the two former species, very short introns (below 50 bp) are relatively rare.

It ought to be mentioned that in many species (mostly animals, including *C.elegans* and *D.melanogaster*) very long introns, albeit infrequent, significantly influence the calculated average values. As a consequence, the average size of introns in *C.elegans* and *D.melanogaster* is 466.6 and 563.9 bp, while the peaks in length distribution lie at ~50 and ~60 bp, respectively (8). In contrast, the distribution of *P.polycephalum* intron sizes is fairly narrow; consequently, the average size (138 bp), the median (98 bp) and the peak in distribution (95 bp) almost coincide. Similar features are seen in *A.nidulans*, whose

introns are significantly shorter (average 72.2 bp; median ~62 bp; peak in distribution ~55 bp) [(8) and our analysis of their figure 2; Kriventseva and Gelfand (15) report slightly higher values].

The lengths of introns and surrounding exons are, across all species, inversely correlated. In fact, it has been shown experimentally, both in *D.melanogaster* as well as in vertebrate cells (12–14), that for efficient splicing the distance between consecutive splice sites has to be 'short' (less than ~500 bp) either across an intron (as is common in non-vertebrates) or across an exon (as often found in vertebrate genes) (26). In *P.polycephalum*, however, we do not find this type of correlation. On the contrary, longer introns (>200 bp) tend to lie between exons longer than average, while the shortest introns are surrounded by exons shorter than average. The significance of this correlation is unknown; it should be stressed here that, with one exception, even the 'long' *P.polycephalum* introns are shorter than the ~500 bp border value defined experimentally.

The distribution of introns within *P.polycephalum* genes is significantly skewed towards the 5'-end. This has been observed for other organisms, in which 3'-most exons tend to be the longest (27). The reason for this non-random distribution is not clear; an explanation based on retroposition has been proposed (28). Another commonly observed clear-cut non-randomness in the distribution of introns concerns their phases: phase zero introns are the most common, and phase two the least common (29–31). In various species the frequencies are: phase zero 39–56%, phase two 20–26%. The average for >11 000 introns in non-homologous genes is phase zero 48%, and phase two 22% (30). In *P.polycephalum* phase zero introns represent 51%, a value well within the range observed for other species, while phase two introns are extremely rare, comprising only 10% of all introns. The reason for such distribution remains a matter of controversy; it has been argued that it reflects the distribution of hypothetical intron insertion sites ['proto-splice sites' (32)] in various genomes, thus giving strong support to the 'introns late' concept (22). However, except for the proto-splice sequence variant GIG in human genes, the correlation between intron phase distribution and the putative proto-splice sites is questionable and certainly cannot, by itself, explain the relative abundance of phase zero introns and the scarcity of phase two introns (31).

From the compilation of nucleotide sequences at exon-intron boundaries shown in Table 1, one can construct consensus sequences for *P.polycephalum*, and, for comparison, for human genes:

P.polycephalum:

5'...A₄₈A₄₈G₆₈I₁₀₀T₁₀₀A₈₄T₆₀G₆₂T₆₄.....T₄₀T₅₉T₄₁Y₈₀A₁₀₀G₁₀₀I₅₄T₄₃...3'

H.sapiens:

5'...M₇₁A₆₁G₈₀I₁₀₀T₁₀₀R₉₅A₆₉G₈₁T₄₅.....Y₈₆Y₈₆NC₇₄A₁₀₀G₁₀₀I₅₃N...3'

The subscripts indicate the percentage frequency of the most common nucleotide at a particular site; M is A or C, R is A or G, Y is C or T. The *P.polycephalum* consensus is not fundamentally different from those found for other organisms (3–7). Minor deviations include position 4 at the 5'-end of intron, which is T in *P.polycephalum* and tends to be A in other species, and position -4 at the 3'-end, which in most species

seems to be random, but shows a preference for T in *P.polycephalum*.

The 3'-end of introns is pyrimidine-rich; however, this enrichment (66% pyrimidines on average for positions -3 to -16) is less pronounced than in some other organisms, most notably humans (6,15), *C.elegans* (4) and Apicomplexa (15). In fact, it is not entirely appropriate to call the 3'-end of *P.polycephalum* introns particularly pyrimidine-rich in the light of the overall preponderance of pyrimidines in *P.polycephalum* introns (65%, see Table 2). Short introns have their 3'-ends barely enriched in pyrimidines (56% on average for positions -3 to -16); this is in fact less than the average for the whole length of these introns (62%, Table 2). Also in other species the preference for pyrimidines at the 3'-end is less marked in short than in long introns (4,12,13). An unexpected feature of long *P.polycephalum* introns is that they never occupy the 5'-most position in their respective genes. In contrast, in several species the 5'-most introns tend to be longer than average (15).

The 5'-ends of short introns are significantly closer to the consensus GTATGT (out of 14 such introns, eight show a perfect match and three a 5/6 match) than the long introns (out of 11, two show perfect match and four a 5/6 match).

There are also other scattered differences between short and long intron sequences (Table 1). Examples include the frequency of Gs at positions -6, -9 and -11 [interestingly, guanines are frequent in short human introns (11)], and adenine and cytosine at position -10.

The exons differ as well: those adjacent to long introns have a much higher than average preference for terminal Gs.

In addition to nucleotide sequences adjacent to the splice sites, another element important in the splicing reaction is the branch site. As there are no experimental data concerning branch site selection in *Physarum*, we decided not to identify such putative sites based solely on sequence analysis [see Kriventseva and Gelfand (15) for a discussion].

The overall nucleotide composition of *P.polycephalum* introns is far from balanced. As found for most other organisms, they are enriched in AT base pairs (61 versus 48% in exons; short introns are slightly more AT-rich, 65%). This degree of enrichment (13%) is similar to that found in *N.crassa* and plants, but significantly less than found in *D.melanogaster*, *Tetrahymena thermophila*, *C.elegans* and *D.discoideum* [21–26%, (3)]. In vertebrate genes this enrichment is negligible.

An unusual feature of *P.polycephalum* introns is the preponderance of pyrimidines in the coding strand. This asymmetry is generally not due to the presence of isolated poly(Py)·poly(Pu) stretches (even though such sequences, occasionally exceeding 50 bp, are found in some introns), but rather is evident throughout the intron length. Except for their 3'-ends (see above), the short and long introns do not differ in this respect. The meaning of this strong and almost universal asymmetry (only four introns out of 64 do not show the preference for CTs in the coding strand), with surrounding exons and intergenic sequences being perfectly balanced (50 and 49% pyrimidines in the coding strand, respectively), is unknown. An interesting observation is that two *P.polycephalum* repetitive, transposon-like sequences, Tp1 (33) and Tp2 (34), also contain regions of highly asymmetric purine/pyrimidine distribution. Short sequences homologous to these putative transposons can be found in some *P.polycephalum* introns (35). These observations

suggest a possible involvement of ancient transposon insertions in the compositional asymmetry of introns.

The number of *P.polycephalum* sequences available for analysis is rather small, which could raise a concern regarding the robustness of the conclusions discussed above. To check the representativeness of the sample analyzed we removed randomly 50% of the sequences, or the sequences belonging to each of the three gene families prominent in the original dataset (actins, tubulins and small GTPases, respectively; see Materials and Methods), and analyzed the remaining sequences. The results (not shown), including the consensus splice site sequences, were not substantially different from those obtained for the whole dataset. One should also note that the sequences analyzed represent several types of independent genes, of diverse expression patterns and intensities, and varied cellular functions. Taken together, these facts indicate that the conclusions presented in this paper should be valid for the whole *P.polycephalum* genome, although minor modifications are likely when more sequences become available.

The data presented in this paper show that the details of exon–intron organization reflect to a certain degree both the evolutionary position and the size of the genome of an organism. However, there are also species-specific features. Only experimental data will tell which of the observed features are functionally relevant, and which are but evolutionary accidents.

These observations also have a direct bearing on the numerous genome sequencing projects currently underway. Care should be taken to train the gene-calling programs on a set of well-characterized genes from the genome being analyzed. Attempts to use programs constructed for other species are likely to lead to erroneous assignments of exon–intron boundaries.

ACKNOWLEDGEMENTS

We thank Dr John Spieth for the information on intronless genes in the *C.elegans* genome. We are also grateful to the anonymous reviewers for helpful suggestions. This work was supported by KBN grant 6 P04A 041 09.

REFERENCES

- Black,D.L. (1995) *RNA*, **1**, 763–771.
- Sharp,P.A. and Burge,C.B. (1997) *Cell*, **91**, 875–879.
- Csank,C., Taylor,F.M. and Martindale,D.W. (1990) *Nucleic Acids Res.*, **18**, 5133–5141.
- Fields,C. (1990) *Nucleic Acids Res.*, **18**, 1509–1512.
- Mount,S.M., Burks,C., Hertz,G., Stormo,G.D., White,O. and Fields,C. (1992) *Nucleic Acids Res.*, **20**, 4255–4262.
- Stephens,R.M. and Schneider,T.D. (1992) *J. Mol. Biol.*, **228**, 1124–1136.
- Conrad,R., Liou,F.N. and Blumenthal,T. (1993) *Nucleic Acids Res.*, **21**, 913–919.
- Deutsch,M. and Long,M. (1999) *Nucleic Acids Res.*, **27**, 3219–3228.
- Goodall,G.J. and Filipowicz,W. (1989) *Cell*, **58**, 473–483.
- Goodall,G.J. and Filipowicz,W. (1991) *EMBO J.*, **10**, 2635–2644.
- McCullough,A.J. and Berget,S.M. (1997) *Mol. Cell. Biol.*, **17**, 4562–4571.
- Guo,M., Lo,P.C.H. and Mount,S.M. (1993) *Mol. Cell. Biol.*, **13**, 1104–1118.
- Talerico,M. and Berget,S.M. (1994) *Mol. Cell. Biol.*, **14**, 3434–3445.
- Sterner,D.A., Carlo,T. and Berget,S.M. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 15081–15085.
- Kriventseva,E.V. and Gelfand,M.S. (1999) *J. Biomol. Struct. Dyn.*, **17**, 281–288.
- Baldauf,S.L. and Doolittle,W.F. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 12007–12012.
- Embley,T.M. and Hirt,R.P. (1998) *Curr. Opin. Genet. Dev.*, **8**, 624–629.
- Mohberg,J. and Rusch,H.P. (1971) *Exp. Cell Res.*, **66**, 305–316.

19. Hardman, N. (1986) *Bioessays*, **5**, 105–111.
20. Mayer, K., Schüller, C., Wambutt, R., Murphy, G., Volckaert, G., Pohl, T., Düsterhöft, A., Stiekema, W., Entian, K.-D., Terryn, N. *et al.* (1999) *Nature*, **402**, 769–777.
21. Ashburner, M., Misra, S., Roote, J., Lewis, S.E., Blazej, R., Davis, T., Doyle, C., Galle, R., George, R., Harris, N. *et al.* (1999) *Genetics*, **153**, 179–219.
22. Logsdon, J.M., Jr (1998) *Curr. Opin. Genet. Dev.*, **8**, 637–648.
23. Bruchez, J.J.P., Eberle, J. and Russo, V.E.A. (1993) *Fungal Genet. Newslett.*, **40**, 89–96.
24. Vinogradov, A.E. (1999) *J. Mol. Evol.*, **49**, 376–384.
25. Elgar, G., Sandford, R., Aparicio, S., Macrae, A., Venkatesh, B. and Brenner, S. (1996) *Trends Genet.*, **12**, 145–150.
26. Berget, S.M. (1995) *J. Biol. Chem.*, **270**, 2411–2414.
27. Hawkins, J.D. (1988) *Nucleic Acids Res.*, **16**, 9893–9908.
28. Boeke, J.D. and Pickeral, O.K. (1999) *Nature*, **398**, 108–111.
29. Fedorov, A., Suboch, G., Bujakov, M. and Fedorova, L. (1992) *Nucleic Acids Res.*, **20**, 2553–2557.
30. Long, M., Rosenberg, C. and Gilbert, W. (1995) *Proc. Natl Acad. Sci. USA*, **92**, 12495–12499.
31. Long, M., de Souza, S.J., Rosenberg, C. and Gilbert, W. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 219–223.
32. Dibb, N.J. and Newman, A.J. (1989) *EMBO J.*, **8**, 2015–2021.
33. McCurrach, K.J., Rothnie, H.M., Hardman, N. and Glover, L.A. (1990) *Curr. Genet.*, **17**, 403–408.
34. Rothnie, H.M., McCurrach, K.J., Glover, L.A. and Hardman, N. (1991) *Nucleic Acids Res.*, **19**, 279–286.
35. Trzcinska-Danielewicz, J. (1999) Ph.D. Thesis. Warsaw University, Poland.